**DS 200 Introduction to Data Sciences**

**Spring 2018 (Yen)**

**A Mini Data Science Project**

**Learning Objectives**

The goal of the project is to provide you with an experience regarding the design of data science project that includes most, if not all, of the following elements:

1. Problem Formulation: The iterative process of formulating a Data Science problem, assessing its feasibility, and refining the problem based on the result of feasibility analysis.
2. Data Gather and Preprocess: Gather relevant data and pre-process them (including tagging a set of data for training a predictive model).
3. Model Construction and Initial Assessment: Construct a Decision Tree-based predictive model using the training data, and perform initial assessment about the quality of the model using (i) confusion matrix by applying the model to a set of labelled data not used for model construction, and (ii) rules extracted from the decision trees.
4. Rigorous Model Assessment: Use cross-validation to rigorously assess the quality of the model.
5. Model Optimization: Systematically adjust the following two parameters in the construction of decision trees and evaluate each parameter setting using cross-validation.
   1. maximum depth of the tree, and
   2. minimum number of instances in a leaf node
6. Model Application: Apply the Decision Tree-based predictive model to new data gathered.

**Two Types of Projects**

You can select your project from the following two types. Due to the different natures of the two projects, not every elements listed above are applicable for Project Type II. Therefore, the required components of the two projects differ slightly.

*Project Type I: Sentiment or Opinion Classification of Twitter Data*

This project builds on labs 1, 3, 4, and 5. The project gathers twitter data on a topic you choose (which has sufficient positive and negative sentiment/opinion tweets for the purpose of building a predictive model). Through inter-tagger analysis for two types of tags (i.e., relevant/irrelevant and positive/negative), you arrived at a set of labelled (i.e., tagged) twitter data that you and your teammate agree. By splitting your tagged data into training set and testing set, you constructed and assessed two decision tree models: (1) The Relevant Tweet Classification Model, and (2) The Sentiment/Opinion Classification Model for X (“X” is the topic you chose for your project). Using data not used for training, you assess the models you constructed using confusion matrix first.

The required elements of Project Type I are A, B, C, and F. (95%)

The optional elements of Project Type I are D and/or E. (5%)

*Project Type II: Breast Cancer Diagnosis Prediction*

This project builds on lab 1 and 6, using two Breast Cancer data sets created by researchers from University of Wisconsin. The first dataset (WDBC) contains 569 instances (357 benign, 212 malignant), 32 attributes (including an ID, diagnosis result, and 30 real-valued input features). The second dataset contains 699 instances (458 benign, 241 malignant). There are 16 instances that contain a missing attribute value.

The required element of Project Type II are C, D, and E for both datasets. (95%)

The optional elements of Project Type II is handling missing values and ranking the degree of importance of features. (5%)

**Project Deliverable #1**

Submit the following information:

* Choice of Project Type
* Names of Team Members

Other information for PD1 depends on your choice of projects:

* If you choose Project Type I, submit the following information

1. Topic of twitter sentiment/opinion classification
2. Keywords and hashtags used to collect tweets
3. Number of tweets gathered and tagged; number of relevant vs irrelevant tweets AFTER inter-tagger analysis; number of positive vs negative tweets AFTER inter-tagger analysis
4. Results of inter-tagger analysis (Kappa measures and Tagging Guidelines) for Relevant/Irrelevant tagging and Positive/Negative tagging
5. A Decision Tree for Relevant Tweet Classification and its Confusion Matrix
6. A complete description of 4 Project Milestones (to be completed as labs 9, 10, 11, and 12).

Grading Rubric: 2, 3, 4, 5: 50%, 6: 50%

* If you choose Project Type II, submit the following information
  1. Jupiter notebook indicating you are able to read at least one of the datasets using pd.
  2. A decision tree for breast cancer diagnosis using one of the datasets.
  3. A complete description of 4 Project Milestones (to be completed as labs 9, 10, 11, and 12).

Grading Rubric: 1,2: 30%, 2: 70%